A COMPARATIVE STUDY OF STATISTICAL METHODS FOR SYNDROMIC DATA ANALYSIS. THE SYNDROMIC SURVEILLANCE SYSTEM FOR THE ATHENS 2004 OLYMPICS

Urania G. Dafni¹², Sotirios Tsiodras², Demosthenes B. Panagiotakos¹, Kassiani Gkolfinopoulou¹², George Kouvatseas¹, Zoi Tsourti¹, George Saroglou²

¹ Laboratory of Biostatistics, Division of Public Health, Department of Nursing, University of Athens, Greece; ² Hellenic Center for Infectious Diseases Control, Ministry of Health, Athens, Greece

Acknowledgements: We would like to thank the members of the Syndromic Surveillance Team, Dimitris Papamihail, Aggeliki Lambrou, Kostas Athanasakis and Ioannis Karagiannis as well as the over fifty health professionals who made every effort to gather quality data from the EDs.

SUMMARY
The Hellenic Center for Infectious Diseases Control (HCIDC) has implemented a syndromic surveillance system based on information from Urgent Care Facilities (Hospital Emergency Departments and Primary Health Care facilities) since July 2002, in the context of the epidemiological surveillance of the 2004 Athens Olympic Games. For detecting outbreaks in syndromic time-series a family of procedures is used, however there is no generally accepted approach. We suggest the use of an algorithm, known as the PULSAR method, based on removing long-term trends from the observed series and identifying peaks in the residual series of surveillance data with cut-offs using a combination of peak height and width. Application in simulated data revealed an effective technique for observing peaks in time-series of syndromic vents. Further validation regarding the correspondence between detected peaks and true biological alerts is needed.

INTRODUCTION
Syndromic time-series are used in the surveillance of natural epidemics or biological attacks. Syndromic surveillance systems were initially developed in the US,
by the Centers for Disease Control and Prevention (CDC) and the New York Department of Health since 1998 and in the wake of the September 2001 events (Mostashari et. al 2003, Das et. al 2003, Ackelsberg et. al 2001). Almost simultaneously other systems were developed, including the ones by the Boston Department of Health (Lazarus et. al 2001), University of Pittsburgh in 1999 (Tsui et. al 2003), the CDC’s drop-in surveillance systems, the Early Notification of Community-based Epidemics (ESSENCE) (Lewis et. al 2002), and others (Zelicoff et. al 2001).

During 2002, the upcoming Athens 2004 Olympics made imminent the need for a ‘real time’ surveillance system that could alert for unexpected communicable disease outbreaks of particular concern during the Olympics, as well as likely clinical presentations of victims of a bioterrorist event. A drop-in syndromic surveillance system similar to the one operating during the Winter 2002 Olympics in Salt Lake city, was established in Greece since July 2002 (Mundorff et. al 2003, Dafni et. al 2003).

Different outbreak detection algorithms have been suggested and are in effect in various operating syndromic surveillance systems especially in the US (Das et. al 2003, Rogerson 2003, Reis et. al 2003, Hutwagner et. al 2003), however, there is no generally accepted procedure for outbreak detection (Lawson 2002). In this paper, a method based on removing long-term trends from the series of observations and identifying peaks in the residual series of data, is presented. An important feature of the proposed approach is that the algorithm raises alerts taking into consideration both height and breadth of signals.

METHODS

Data Acquisition. Drop-in syndromic surveillance in Emergency Departments (EDs) of major hospitals was first established in Greece by the Hellenic Center for Infectious Diseases Control (HCIDC), in July 2002. From August 2002 through August 2003, the syndromic surveillance system operated in eight major hospitals and a major health care center in the greater Athens area.

The syndromic surveillance included ten syndromes, however in the work presented here, the time series for the two most commonly encountered syndromes, ‘respiratory infection with fever’ and ‘gastroenteritis (diarrhoea, vomiting), without blood’, are used.

Algorithm description. The method evaluated here, known as the PULSAR method, is based on identifying peaks in the syndromic time series, exceeding a specified threshold. Long-term changes are first screened out and then peaks are identified in the screened series. This approach was suggested by Merriam and Wachter for the study of episodic hormone secretion (Merriam and Wachter 1982).

In the first step of the procedure, a baseline is defined for the original syndromic series through the LOWESS smoother (Cleveland 1979) using the bias corrected
Akaike’s Information Criterion (AIC) that incorporates both the tightness of the fit and the model complexity. This criterion often selects better models than AIC in small samples (Hurvich and Simonoff 1998). This procedure produces baseline estimates that are not influenced by extreme outlier observations.

A residual series, containing short-term variations but not the trends, is obtained by subtracting the smoothed data from the original counts. The residual series is standardized by dividing the residuals by an estimate of the noise level, to yield a scaled residual series, expressed in ‘signal-to-noise’ units. The peaks in the standardized residual series are identified based on some combination of height and width, with no assumption for the shape of the peak. To classify an elevation as a peak it should either be very high even if narrow or it should span several points in width even if moderately high. In order for a point in the ‘signal-to-noise’ series to be considered part of a peak, it should exceed a certain cut-off value G(1); or it should exceed a lower cut-off value G(2) along with one adjacent point; or an even lower cut-off value G(3) along with two adjacent points; and so on. After the initial identification of peaks, the baseline is recalculated. Reduced weight is now assigned to observations previously identified as part of a peak. Iterations of the whole process are performed until the same assignment of points to peaks is achieved.

**Algorithm Customization.** Six alternative estimates were used for the standardization, including the standard deviation and the mean absolute deviation in the original series, as well as the 7-day moving standard deviation and the 7-day mean absolute deviation in the simulated series. The latter were based either on the seven most recent observations to the current time point or on the tenth to fourth most recent observations, i.e., not taking into account the three most recent ones. The procedure is performed iteratively so as to down-weigh extreme values and to be able to detect outliers that appear in clusters. In our series, G(1), G(2) and G(3) cut-offs were chosen under the assumption of normality for the standardized residual series, so as to derive 97% specificity in the whole series and take into account the effect of multiple testing on the significance level. The threshold is given by $G(n) = \text{probit}(1 - \alpha \cdot (n/6)/d)$ where $d =$ number of days that a false alarm occurs with probability $\alpha = 0.10$ and $n = 1$ or 2 or 3, while the factor $n/6$ provides the necessary adjustment for multiple testing ($\text{probit}$ stands for the inverse Normal density function).

**Alternative Methods.** In order to evaluate the performance of the PULSAR approach, we compared it through simulations, to other commonly used methods in syndromic surveillance (Rogerson 2003, Reis et. al 2003, Hutwagner et. al 2003, Hutwagner 1997). These methods were: a) The Temporal Aberration Detection (TAD) approach used in the Early Aberration Reporting System (EARS) (Hutwagner et. al 2003), b) time series
methods such as Auto-Regressive Integrated Moving Average (ARIMA) (Reis et al. 2003) and c) the CUSUM (Rogerson 2003, Hutwagner et. al 1997).

All parameters for each model used in the comparisons were set such that the specificity (true non-alarms / non-outbreaks) in the original time series was fixed at 97%, assuming no-outbreak condition (Reis et. al 2003). For each method, the day of an outbreak on which an alert was raised for the first time was recorded. Sensitivity (true alarms / outbreaks) across all simulated series for each syndrome, and the ‘timeliness’ for each method, i.e., the percentage of the first alert per day of outbreak, were compared among the alternative approaches. The three performance criteria, i.e., the sensitivity, specificity and timeliness of the methods are reported and compared through the Wilcoxon signed-rank or Friedman non-parametric tests. Bonferroni adjusted $\alpha^*$ are reported. The methods mentioned below have been used in syndromic surveillance and were evaluated in our syndromic data series.

Simulation Schema. To evaluate the performance of the proposed methodology, 100 simulated series were created. Our original time series of counts is considered to include no outbreaks. A four-day outbreak was chosen to represent a probable time period between symptom presentation and diagnosis, i.e., the window of opportunity (Mostashari 2002). Each simulated time series was produced randomly injecting four-day long outbreaks to the original time series of daily counts for each syndrome of interest with probability 15% per day. An outbreak led to duplication of the observed counts of the syndrome for that day. Two adjacent outbreaks were forced to be at least 15 days apart to ensure that a previous outbreak did not adversely affect the alert detection mechanism of the next (Reis et. al 2003). An outbreak was considered as successfully detected, if an alarm was raised at least one of the days of the outbreak. Alternative patterns of outbreaks were also examined, nevertheless details are not reported here. All statistical computations were performed using SAS 8.2 software (SAS Institute, Inc., Cary, NC, USA).

RESULTS

The original thirteen-month time series for four major hospitals in metropolitan Athens sharing the same catchment area, for the ‘Respiratory Infection with Fever’ syndrome and the ‘Gastroenteritis (diarrhoea, vomiting), without Blood’ syndrome are used here for illustrating and evaluating the proposed method.

First, the six different standardization estimates already described for the PULSAR algorithm, leading to different threshold specifications, were compared and the best approach for both syndromes with respect to the achieved sensitivity was the one using the standard deviation in the original series. The TAD approach was used both on the count series and on the proportion of counts of syndromes to total ED visits (Das et.al 2003, Hutwagner et. al 2003). In our data, the results for the count series were
superior to the ones for the proportion series. Regarding the ARIMA models the best filter regarding sensitivity was the 7th order MA filter, while the weekend was also statistically significant and was used as an explanatory variable in the model. For the one-sided CUSUM method used here, a 7-day moving average and standard deviation used for standardization proved superior to the standard approach. A second approach using values commonly used in bibliography that actually minimize the Average Run Length (ARL) of the process, were also used (k=0.5 and h=2.5) (Rogerson 2003).

The PULSAR approach fairs well in comparison to the other methods for each evaluation criterion. In particular, mean sensitivity was statistically significantly higher (Bonferroni $\alpha^*=0.0056$) for the PULSAR approach when compared to the other approaches for both syndromes (Wilcoxon signed-rank p-value $<0.001$ for all comparisons). Furthermore, mean specificity for the PULSAR method was significantly higher (Bonferroni $\alpha^*=0.0056$) than the specificity of the one-sided CUSUM method (Wilcoxon signed-rank p-value $<0.001$). This finding holds for both syndromes examined. In addition, in the case of 'Respiratory Infection with Fever', the specificity of PULSAR is significantly higher than the specificity of TAD (Wilcoxon signed-rank p-value $<0.001$), while in the case of 'Gastroenteritis', the specificity of PULSAR is higher than the specificity of the ARIMA approach (Wilcoxon signed-rank p-value $<0.001$). No other significant differences regarding specificity between the PULSAR method and the other ones were found for either syndrome. Timeliness for the first day (proportion of alerts at the first day of an outbreak) differed significantly among the four approaches (Friedman test, p-value $<0.001$) for both syndromes. Timeliness for the PULSAR method is lower than the timeliness of the ARIMA model (Wilcoxon signed-rank p-value $<0.001$, Bonferroni $\alpha^*=0.0056$). However, in the case of 'Respiratory Infection with Fever', PULSAR’s timeliness is higher than the one of TAD and CUSUM approaches (p-values $<0.001$), while in the case of 'Gastroenteritis', PULSAR’s timeliness is higher than TAD’s (p-value $<0.001$). Results were similar when using the mentioned alternative patterns of outbreaks.

DISCUSSION

The PULSAR approach, first suggested for the study of episodic hormone secretion, was successfully used in the context of syndromic surveillance data. Initially, syndromic data are expressed in ‘signal-to-noise’ units, and then through an iterative process, peaks are identified. Point elevations that are very high or elevations only moderately high but spanning several points in width, are identified as peaks. The thresholds for peak detection are probabilistically determined based on the assumption of normally distributed residuals. The idea of stochastically determining the thresholds is extended to the other methods under comparison. The thresholds are chosen such that a specificity of 97% is achieved in the original syndromic time-series.
In the simulated datasets, the 97% specificity was most closely reproduced in the case of the proposed method as compared to the others. Sensitivity for the chosen PULSAR model for the ‘Respiratory Infection with Fever’ syndromic series, ranged from 67% to 100%, with a mean of 85% while for the ‘Gastroenteritis (diarrhoea, vomit) without Blood’ syndrome, ranged from 62.5% to 100%, with a mean of 81%. The mean sensitivity for the PULSAR approach was higher than the sensitivity for the other methods. Our method compared well with the others as far as specificity. All methods held specificity close to the 97% benchmark with the exception of the one-sided CUSUM. In all methods evaluated, the higher percentage of alerts was raised on the first day of the outbreak with the exception of the TAD model where the alerts occur with similar frequency between the first three days of the outbreak. The ARIMA model exhibited the best timeliness results followed by the PULSAR approach.

It is noted that the performance evaluation criteria led to uniformly worse results for all methods when applied to the daily proportion of syndrome to total visits as opposed to counts. Another limitation is that a specific simulation schema was used for the comparison of methods, a fact that affects the generalization of the comparison under other simulation settings. However, the critical comparison is always the one based on the detection performance of real outbreaks (Sosin 2003). Moreover, a variety of methods have been proposed but were not considered here (Sonesson 2003). These include spatial statistical methods like the spatial scan statistic, Bayesian approaches and multivariate methods.

The performance results of the PULSAR method are overall comparable to the other methods examined for the specific simulation schema used. In addition, its ability to detect peaks based not only on height but also on width, addresses more closely the epidemic shapes that one would expect to last for more than one day. The Athens 2004 Olympic Games, a unique experiment due to the expected abrupt increase in population, could provide the ideal prospective surveillance setting for comparing the behaviour of all proposed methods regarding alert mechanisms.

ΠΕΡΙΛΗΨΗ

Εισαγωγή. Δεν υπάρχει μια γενικώς αποδεκτή διαδικασία για την ανίχνευση εξάρσεων σε χρονοσειρές συνδρομικών δεδομένων στα πλαίσια της επιτήρησης φυσικών επιδημιών ή βιοτροµοκρατικών επιθέσεων. Σκοπός. Προτείνεται η χρήση μιας προσέγγισης γνωστής ως µέθοδος PULSAR, η οποία βασίζεται στην αποµάκρυνση των µακροπρόθεσµων τάσεων από τις παρατηρούµενες χρονοσειρές και την εντόπιση κορυφών στις κατάλοιπες σειρές των συνδροµικών δεδοµένων, µέσω ορίων τα οποία προκύπτουν από το συνδυασµό του ύψους µε την χρονική διάρκεια της κορυφής. Μέθοδος. Πραγµατοποιήθηκαν προσοµοιώσεις για την αξιολόγηση της µεθόδου και τη σύγκριση της µε εναλλακτικές προσέγγισες. Αναλύθηκε το ηµερήσιο πλήθος συνδροµικών παρατηρήσεων στα Τµήµατα Επειγόντων Περιστατικών (ΤΕΠ) τεσσάρων µεγάλων νοσοκοµείων της περιοχής της Αθήνας για τα δύο συχνότερα σύνδροµα από την 1η Αυγούστου 2002 έως την 31η Αυγούστου 2003 (Πηγή: Σύστηµα
Συνδρομικής Επιτήρησης Ο.Α. 2004). Παράχθηκαν τυποποιηµένες σειρές καταλοίπων μετά την αφαίρεση της τάσης και του θορύβου από τις αρχικές σειρές. Οι σειρές εξετάστηκαν για την παρουσία κορυφών, οριζόµενες ως σηµεία τα οποία ξεπερνούσαν σε ύψος τουλάχιστον ένα από τρία πιθανοθεωρητικά προσδιορισµένα ανώτατα όρια. Επαναλήψεις της διαδικασίας επαναπροσδιόριζαν την χρονοσειρά αναφοράς, αναθέτοντας μειωµένα βάρη στις κορυφές που εντοπίζονταν. **Αποτελέσµατα.** Συγκρινόµενη με τις υπόλοιπες μεθόδους, και για το συγκεκριµένο σχήµα προσοµοίωσης, η µέθοδος PULSAR ανταποκρίθηκε ικανοποιητικά αναφορικά µε τα κριτήρια της ευαισθησίας, ειδικότητας και χρόνου ανταπόκρισης. **Συµµετάσχοντες.** Ο προτεινόµενος αλγόριθµος χρειάζεται περαιτέρω εξέταση βάσει της αντιστοιχίας διεγνωσµένων κορυφών και πραγµατικών βιολογικών συναγερµών. Παρόλο πολύ η εφαρµογή σε προσοµοιοµένα δεδοµένα απεκάλυψε µια αποτελεσµατική τεχνική για τον εντοπισµό κορυφών σε χρονοσειρές συνδροµικών περιστατικών. Η απλότητα του αλγορίθµου, η ικανότητα του να εντοπίζει κορυφές, τόσο σε σχέση µε το ύψος όσο και τη διάρκεια, και η συµπεριφορά του στα προσοµοιοµένα δεδοµένα τον καθιστούν έναν παραιτέρω χρήση στο πλαίσιο της συνδροµικής επιτήρησης.

**REFERENCES**


